BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

09869312

PCT/JP00/00355

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

25.01.00

4

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

JP00/311

出願年月日
Date of Application: １９９９年 １月２５日

出願番号
Application Number: 平成１１年特許願第０１５１８９号

出願人
Applicant (s): 株式会社医薬分子設計研究所

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

２０００年 ２月２５日

特許庁長官
Commissioner,
Patent Office

近藤隆彦

出証番号 出証特２０００－３００９５５８

【書類名】　特許願

【整理番号】　98165M

【提出日】　平成11年 1月25日

【あて先】　特許庁長官　殿

【発明者】

【住所又は居所】　東京都武蔵村山市三ツ藤１－６５－５

【氏名】　豊田　哲郎

【発明者】

【住所又は居所】　東京都文京区本郷５－１６－６

【氏名】　板井　昭子

【特許出願人】

【識別番号】　597051148

【氏名又は名称】　株式会社医薬分子設計研究所

【代理人】

【識別番号】　100096219

【弁理士】

【氏名又は名称】　今村　正純

【選任した代理人】

【識別番号】　100092635

【弁理士】

【氏名又は名称】　塩澤　寿夫

【選任した代理人】

【識別番号】　100095843

【弁理士】

【氏名又は名称】　釜田　淳爾

【手数料の表示】

【予納台帳番号】　038357

【納付金額】　21,000円

出証特２０００－３００９５５８

【提出物件の目録】

【物件名】　明細書　1

【物件名】　要約書　1

【包括委任状番号】　9707349

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　アラインメント情報の保存方法

【特許請求の範囲】

【請求項１】　アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離することを特徴とする、アラインメント情報の記述方法。

【請求項２】　アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離し、これらの情報のうち少なくともギャップ情報を記録媒体に保存することを特徴とする、アラインメント情報の保存方法。

【請求項３】　ギャップ情報のみを保存する、請求項２に記載の方法。

【請求項４】　ギャップ情報が、２以上の配列間のアラインメント情報に存在するギャップ部分の位置及び長さを示す残基番号及び／又は残基数のデータ、又は該データに計算変換可能な数値データに基づいて記述されたものである、請求項１ないし３のいずれか１項に記載の方法。

【請求項５】　ギャップ情報が、アラインメントに含まれない他の配列または仮想の配列の残基番号を含むデータ又は該データに計算変換可能な数値データに基づいて記述されたものである、請求項１から３のいずれか１項に記載の方法。

【請求項６】　アミノ酸配列又は核酸配列のアラインメント情報から配列情報を含まないように分離された配列間の対応関係を表すギャップ情報と、対応する配列情報とから、通常の表現形式のアラインメント情報を得る方法。

【請求項７】　アミノ酸配列又は核酸配列のアラインメント情報から配列情報を含まないように分離された配列間の対応関係を表すギャップ情報のみに基づいて、新たな配列間のアラインメント情報についてのギャップ情報を演算により生成させる方法。

【請求項８】　請求項７に記載の方法に従って生成させたギャップ情報と対応の配列情報とを用いて、通常の表現形式のアラインメント情報を得る方法。

【請求項９】　アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離し、これらの情報のうち少なくとも

ギャップ情報を通信することを特徴とする、アラインメント情報の通信方法。

【請求項１０】　複数のアラインメント情報の通信において、それぞれのアラインメント情報を配列情報とギャップ情報とに分離し、配列情報ごとおよびギャップ情報ごとにまとめて通信する方法。

【請求項１１】　複数のアラインメント情報から分離された配列情報における重複を除いて通信する請求項10に記載の方法。

【請求項１２】　アラインメント情報のデータベースであって、請求項２ないし４のいずれか１項に記載の方法により少なくともギャップ情報を記録媒体に保存したデータベース。

【請求項１３】　ギャップ情報に替えて対応付け情報を用いる、請求項１ないし１１のいずれか１項に記載の方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、複数のアミノ酸配列間でアミノ酸残基、又は複数の塩基配列間で核酸残基をできる限り一致するように対応付けて並置したアラインメント情報を、通常の表現形式に比べてより少ないデータ量で保存又は通信する方法に関するものである。

【０００２】

遺伝情報を担う核酸の塩基配列の情報は、アミノ酸配列に翻訳される。蛋白質の多様な機能や立体構造は２０種のアミノ酸残基の並び方（アミノ酸配列）によって決定されるが、アミノ酸配列の情報からその蛋白質の機能及び立体構造を直接推定することは困難であり、その解明には多大な実験的労力が必要である。近年、ゲノム情報の解析が進んだ結果、蛋白質を単離することなく、生体に存在している蛋白質のアミノ酸配列が解明されるようになった。このため、アミノ酸配列の情報から蛋白質の機能及び立体構造を研究するための手法がますます重要になっている。

【０００３】

一般に、アミノ酸配列の類似度が高い複数の蛋白質間では機能や立体構造が類似

している確率が高いという経験則に基づいて、間接的に蛋白質の機能や立体構造を推定することができる。また、同じ機能をもつ蛋白質でも生物種や組織によって配列が相当に異なっており、その類似の具合から進化的な関係を推論することもよく行われる。そうした場合、複数の配列間でアミノ酸残基ができるだけ対応するよう、対応付けをする作業（アラインメントという）が行われる。アラインメント(alignment)とは、複数配列間でアミノ酸の対応関係を探しだすこと、及びその結果である対応関係を指す。対応付けはアミノ酸残基間の同一性や相同性、及び立体的位置関係の近さなど様々な指標によって評価される（以下、アミノ酸配列のアラインメントについて具体的に説明するが、本明細書において用いられる用語「アラインメント」には、塩基配列のアラインメントも含まれる。）。アラインメント情報は対応づけられたアミノ酸残基同士が配列間で並ぶように文字列で表現される（総説として、美宅成樹及び金久實著、「３．文字を比較し並べる」、培風館、1995年）。

【０００４】

アラインメントは配列情報を科学研究や産業に利用するための手段であり、アラインメントを必要とする研究が増加している。例えば、同じ機能の複数蛋白質を複数の生物種間で比較して機能と構造との関係を調べたり、結晶解析するかわりに立体構造既知の類縁蛋白質の結晶構造に基づいてモデリングするなど、既知蛋白質間でもアラインメントは頻繁に行われている。その結果、アラインメント結果は生化学、分子生物学、遺伝子工学などの分野の学術論文誌に頻繁に記載されている。しかしながら、一般的には、アラインメント情報は使い捨てにされており、各研究者が、その都度必要な蛋白質群の配列からアラインメントを作成しているのが現状である。アラインメント情報をデータベース化することにより、他の研究への応用が可能になり、研究の効率化が進むことが期待される。また、アラインメントの手法はほぼ標準化されているところから、ゲノム解析から新たに存在が知られる蛋白質について、標準的な処理によって既知蛋白質とアラインメントし、その情報を上記のデータベースに加えることも有用と思われる。

【０００５】

もっとも、アミノ酸配列にハイフンを加えてギャップを表した通常の表現形式の

アラインメント情報は、研究者が視覚的に類似性を理解するには都合がよいものの、“（配列の残基数＋ギャップの残基数）×配列数”の数の文字を記憶させる必要があることから、膨大なアラインメント情報を保存して再利用するためには適しない。文字のまま保存することは情報処理の観点からは無駄が多く、また、配列情報自体は一般に配列情報データベースから取得できるので、重複した情報を保存することにもなる。さらに、新しいアミノ酸配列が加わったり、配列の一部が削除されてギャップの位置や長さが変わった場合には、アラインメントの更新が必要になる。

【０００６】

今後、アミノ酸配列の情報が加速度的に増加し、アラインメント情報の利用も急増することが予想されることから、アラインメント情報をコンピュータ上に効率的に保存し、検索や編集を可能にする方法の開発が求められている。また、現在はコンピュータのネットワーク化が進み、アラインメント情報のデータも複数のコンピュータ間で頻繁に送受信されるため、アラインメント情報の効率的な通信方法が必要である。

【０００７】

【発明が解決しようとする課題】

本発明の課題は、アラインメント情報を効率的に保存し、又は通信する方法を提供することにある。より具体的には、アラインメント情報を少ないデータ量で保存又は通信することができ、検索や編集が可能で、かつ必要時には通常の表現形式のアラインメント情報を迅速に取り出すことができる方法を提供することが本発明の課題である。また、本発明の別の課題は、上記の方法で保存されたアラインメント情報を格納した記録媒体又はデータベースを提供することにある。さらに別の課題は、上記の方法でデータの通信を行うためのプロトコルを提供することにある。

【０００８】

【課題を解決するための手段】

アラインメント情報は、通常、一文字表記で表したアミノ酸残基を「残基番号」（各配列でＮ末端から数えたアミノ酸残基の順番をいう）の順に左から右へ列記

した複数のアミノ酸配列を縦に重ね、対応付けたアミノ酸残基を同じ欄（縦列方向）に置くことにより各アミノ酸配列間の対応関係を表現した情報として提示される（表１）。以下、本明細書において、アラインメント情報のこの表現形式を「通常の表現形式」と呼ぶ。それぞれの配列に含まれるアミノ酸残基のうち、同じ欄に置かれたアミノ酸残基は互いに対応付けられたことを意味しており、いずれか一方の配列に対応する残基がない場合はハイフンを挿入して表現される。この様なハイフン（又はハイフンのつながり）は「ギャップ」と呼ばれる。「ギャップ」でない部分はすべて対応付けられている。

【０００９】

【表１】

アミノ酸配列のアラインメント情報

配列Ａ：　--MISLIAALAVD-VIMGRHTWESIVYEQFLPKAQHDLYIA--

配列Ｂ：　RSMLSIVAVCQNDAVIMGKKTWFSIVY----AKAQHEKFVSPA

【００１０】

本発明者らは、アラインメント情報が、「配列情報」と「ギャップ情報」（各ギャップが挿入される残基番号とギャップ数、又は各ギャップ部分に対応する他方の配列の残基番号と残基数）又は「対応付け情報」（各対応付けされた部分の残基番号と残基数を示す情報）に分離できることに着目した。「配列情報」は20種類のアミノ酸残基又は４種類の核酸残基を区別するための文字情報である。一方、「ギャップ情報」又は「対応付け情報」は残基番号又は残基数で表される数値データであり、両者は等価な情報として相互に変換可能である。本発明者らは、分離されたギャップ情報（又は対応付け情報）を配列情報と組み合わせることにより、容易に「通常の表現形式」のアラインメント情報に変換することができることを見出した。

【００１１】

一般的に、配列中のギャップ部分の数は一般にアミノ酸残基数に比べると少なく、概ね10分の１以下であるところから、ギャップ情報（又は対応付け情報）と配列情報とをアラインメント情報から分離し、ギャップ情報（又は対応付け情報）

のみを保存又は送信することにより、アラインメント情報を極めて少ないデータ量で効率的に保存又は通信できる。また、一般に、配列情報は他の利用可能な配列データベースから得られることが多いので、アラインメント情報を保存又は通信するためには、ギャップ情報（又は対応付け情報）のみを取り扱えばよい。一方、配列情報が容易に利用可能でない場合には、ギャップ情報とともに配列情報を保存又は通信することもできる。通常の表現形式のアラインメント情報には配列情報が含まれているため、この形式のアラインメント情報と配列情報とを保存又は通信すると、配列情報を重複して取り扱うことになる。ギャップ情報と配列情報とを分離して取り扱うことによって、このような重複がなくなり、保存又は通信の効率化が期待できる。本発明はこれらの知見を基にして完成された。

【００１２】

すなわち本発明は、アラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離して保存又は通信する方法、及びアラインメント情報を配列情報と配列間の対応関係を表す対応付け情報とに分離して保存又は通信する方法を提供するものである。配列情報が既存のデータベースから利用できる場合には、アラインメント情報を配列情報とギャップ情報又は対応付け情報とに分離し、ギャップ情報又は対応付け情報のみを保存又は通信することが可能である。上記ギャップ情報又は対応付け情報は、配列情報を含まず、少数の数値データで表現できるうえ、配列情報を用いた計算処理により通常の表現形式のアラインメント情報に変換することができる。

【００１３】

本発明により、下記の方法が提供される。

(1)アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離することを特徴とする、アラインメント情報の記述方法；

(2)アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離し、これらの情報のうち少なくともギャップ情報を記録媒体に保存することを特徴とする、アラインメント情報の保存方法。

(3)ギャップ情報のみを保存する上記(2)の方法；

(4)ギャップ情報が、２以上の配列間のアラインメント情報に存在するギャップ部分の位置及び長さを示す残基番号及び／又は残基数のデータ、又は該データに計算変換可能な数値データに基づいて記述されたものである、上記(1)ないし(3)のいずれかの方法；

【００１４】

(5)ギャップ情報が、アラインメントに含まれない他の配列または仮想の配列の残基番号を含むデータ又は該データに計算変換可能な数値データに基づいて記述されたものである、上記(1)から(3)のいずれかに記載の方法；

(6)アミノ酸配列又は核酸配列のアラインメント情報から配列情報を含まないように分離された配列間の対応関係を表すギャップ情報と、対応する配列情報とから、通常の表現形式のアラインメント情報を得る方法；

(7)アミノ酸配列又は核酸配列のアラインメント情報から配列情報を含まないように分離された配列間の対応関係を表すギャップ情報のみに基づいて、新たな配列間のアラインメント情報についてのギャップ情報を演算により生成させる方法；

(8)上記７に記載の方法に従って生成させたギャップ情報と対応の配列情報とを用いて、通常の表現形式のアラインメント情報を得る方法；

【００１５】

(9)アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離し、これらの情報のうち少なくともギャップ情報を通信することを特徴とする、アラインメント情報の通信方法；

(10)複数のアラインメント情報の通信において、それぞれのアラインメント情報を配列情報とギャップ情報とに分離し、配列情報ごとおよびギャップ情報ごとにまとめて通信する方法；

(11)複数のアラインメント情報から分離された配列情報における重複を除いて通信する(10)に記載の方法；及び

(12)ギャップ情報に替えて対応付け情報を用いる、上記(1)ないし(11)のいずれかの方法。

【００１６】

また、別の観点からは、アラインメント情報の保存方法であって、上記ギャップ情報又は対応付け情報をデータベース又は記録媒体中に保存する方法；上記ギャップ情報又は対応付け情報を配列情報とともにデータベース又は記録媒体中に保存する上記方法；及び、ギャップ情報又は対応付け情報を同一又は分散化された形態のデータベース又は記録媒体中に保存する上記方法が提供される。

【００１７】

さらに別の観点からは、アラインメント情報を再現するために最小限必要な情報のみを通信するための方法が提供される。すなわち、アラインメント情報の通信方法であって、上記ギャップ情報又は対応付け情報を通信する方法；配列情報と上記ギャップ情報又は対応付け情報とを通信する上記方法；及び、配列情報とギャップ情報又は対応付け情報とを同一又は分散化された形態で通信する上記方法が提供される。

【００１８】

これらの発明に加えて、ギャップ情報又は対応付け情報を含むアラインメント情報のデータベース；配列情報とギャップ情報又は対応付け情報とを含むアラインメント情報のデータベース；配列情報とギャップ情報又は対応付け情報とに分離されたアラインメント情報を含むデータベース；ギャップ情報又は対応付け情報を格納した記録媒体；配列情報とギャップ情報又は対応付け情報とを格納した記録媒体；及び、配列情報とギャップ情報又は対応付け情報とに分離されたアラインメント情報を含む記録媒体が提供される。記録媒体の種類は特に限定されず、当業者に利用可能な光ディスク、磁気ディスク、磁気テープなどを用いることができる。また、本発明により、配列情報とギャップ情報又は対応付け情報とに分離されたアラインメント情報を通信するためのプロトコルが提供される。

【００１９】

【発明の実施の形態】

以下、本発明の方法を２つのアミノ酸配列から得られたアラインメント情報に対して適用する場合について具体的に説明するが、本発明の範囲は下記の態様及びその説明の細部に限定されることはない。また、以下の説明においては、アラインメント情報を配列情報とギャップ情報とに分離する場合についてのみ言及する

が、「ギャップ情報」と「対応付け情報」とは等価であり、互いに変換可能であることから、本発明の方法がギャップ情報を用いる方法のみに限定されると解釈してはならない。

【００２０】

さらに、本明細書において用いられる「配列」と言う用語は、特に言及しない場合には、アミノ酸配列及び核酸配列のいずれをも含む概念として用いる。また、本明細書において用いられる「アラインメント情報」という用語は、論文誌上で公開されるアラインメント結果や標準的手法で得られる通常の表現形式で記述されたアラインメント結果のほか、通常の表現形式以外で表現されたアラインメント結果、種々の解析方法の実行にあたり中間データとして生成されるアラインメント結果、及びアラインメント結果の部分的情報などを含めて、最も広義に解釈する必要がある。また、本明細書において用いられる「保存」という用語は、ギャップ情報などを記録媒体又はデータベース等に保存する作業のほか、保存されているギャップ情報を統合したり、保存されているギャップ情報から通常の表現形式のアラインメント情報を再現するなど、保存された情報の利用を含む概念として用いる。

【００２１】

上記表１に示した２つの配列Ａ及びＢのアラインメント情報から配列情報を除くと表２に示した情報が得られる。数字は残基番号を示しており、アミノ酸配列の情報から各残基番号に対応するアミノ酸の種類がわかるので、本発明の方法では、表２においてハイフンで表されるギャップ部分の位置と長さのみを別の形式で記述して保存する。

【００２２】

【表２】

A	－	－	1	2	3	4	5	6	7	8	9	10	11	－	12	13	14	15	16	17	18
B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

A	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	－
B	22	23	24	25	26	27	－	－	－	－	28	29	30	31	32	33	34	35	36	37

【００２３】

本発明の方法は、表２に示したような配列情報を含まないアラインメント情報を少数の数字データを用いて保存することを特徴としており、その手段として、配列間の対応関係を表す「ギャップ情報」を利用する。以下、アラインメント情報からギャップ情報を抽出する手法の具体例を説明するが、ギャップ情報の種類又はその記述方法は以下に説明するものに限定されることはない。ギャップ情報としては、配列間の対応関係を表すものであれば、いかなるものを利用してもよい。なお、以下の説明において、各配列の残基数の情報は、配列情報に含まれているものとする。

【００２４】

＜ギャップ情報の記述法＞

あるアラインメント中の各配列に存在するギャップの位置及び長さは、ギャップ部分の位置を示す残基番号や含まれる残基数によって記述できる。記述の方法は特に限定されず、いかなる方法を採用してもよい。以下に代表例を示すが、記述方法はこれらに限定されることはない。アラインメントは複数配列間の相対的な関係を示すものであり、基準とする配列に基づいて記述するか否か、配列をどう選ぶかなどの要素によって、下記の方法①、②、③に大別できる。方法①は、アラインメントに含まれる複数の配列のうちの一の配列の残基番号に基づいて記述する方法である。方法②では、通常の表現形式で記述するときの欄番号、または現実のアミノ酸配列ではない仮想配列の残基番号その他に基いて記述する。方法③では、アラインメント中の各配列につき、ギャップ部分と残基の存在する部分の残基数を交互に並べることで、基準とする配列に基づかずに、ギャップ情報を

記述する。さらにこれらの各方法においても、さまざまな修飾や改変が可能である。なお、残基数のみでギャップ情報を表現する場合、残基番号の小さいギャップから順に記述することが望ましい。

【００２５】

Ａ配列を基準配列とし、表２のアラインメント情報をギャップ部分と対応付けされた部分の各残基数を交互に（ギャップ部分の残基数を先に）並べると「２，１１，１，１３，－４，９，１」と記述できる（方法①ａ）。ギャップ部分がどの配列に存在するかを示すため、Ｂ配列中のギャップの残基数はマイナスの符号を付して負の数としている。この記述は、左端から、Ａ配列に２残基のギャップ、１１残基の対応付け部分、Ａ配列に１残基のギャップ、１３残基の対応付け部分、Ｂ配列に４残基のギャップ、９残基の対応付け部分、最後にＡ配列に１残基のギャップがあることを意味している。

【００２６】

また、表２のアラインメントは、残基番号と残基数を用いて（残基番号を先に書いた場合）、「０，２，１１，１，２４，－４，３７，１」と記述することもできる（方法①ｂ）。この記述では、各ギャップをその直前の残基番号と残基数で示し、Ｂ配列のギャップは負の数とすることでＡ配列と区別する。この記述は、Ａ配列の残基番号０番（Ｎ末端）に２残基のギャップ、Ａ配列の１１番の後に１残基のギャップ、Ａ配列の残基番号で２４番に対応する位置の後からＢ配列に４残基のギャップ、Ａ配列の３７番の後ろに１残基のギャップと並ぶことを意味している。

【００２７】

方法①ａと方法①ｂの記述は、方法①ａで対応付けされた部分の残基数を加えて残基番号とすれば方法①ｂに変換できる。方法①ａの「２，１１，１，１３，－４，９，１」から、先頭のギャップは０番から始まることにして０をおき、１１はそのまま、１３は２４（＝１１＋１３）、９は３７（＝２４＋４＋９）（４はＢ配列のギャップで、Ａ配列には残基が存在するので）とすることにより、「０，２，１１，１，２４，－４，３７，１」に変換できる。逆の手順により、逆の変換も可能である。

【００２８】

もっとも、表３のような３以上の配列を含むアラインメント情報を記述し、通常の表現形式を再現したり、アラインメント中の配列の増減を行う目的には、全配列を対等に扱う方法が便利である。以下、３以上の配列のアラインメントの例（表３）を用いて、全配列を対等に扱う方法②と③について説明し、さらに通常の表現形式への変換の方法、配列の削除、アラインメントの統合について説明する。

【００２９】

【表３】

A	－	－	1	2	3	4	5	6	7	8	9	10	11	－	12	13	14	15	16	17	18	19	20
B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
C	－	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

A	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	－				
B	24	25	26	27	－	－	－	－	28	29	30	31	32	33	34	35	36	37				
C	23	24	25	－	－	－	26	27	28	29	30	31	32	33	34	35	36	－				

【００３０】

ほとんどのアラインメント情報において、ギャップ部分の存在のため、残基の対応関係を示す欄の数は最も長い配列の残基数より長いのが普通であり、どの配列の残基番号とも一致しない。表３のアラインメント情報に対して、残基またはハイフンを置く欄に通し番号を付したのが表４である（欄の行はＲと記してある）。

【００３１】

【表４】

R	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	－	－	1	2	3	4	5	6	7	8	9	10	11	－	12	13	14	15	16	17	18	19
B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
C	－	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

R	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
A	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	－
B	23	24	25	26	27	－	－	－	－	28	29	30	31	32	33	34	35	36	37
C	22	23	24	25	－	－	－	26	27	28	29	30	31	32	33	34	35	36	－

【００３２】

方法②

こうして得られた欄の番号に基づいて（Ｒを基準配列として）、各配列につき、残基の存在する部分（またはギャップ部分）の位置を記述するのが方法②である。この方法では、それぞれの配列の間での対応付けは必要ない。この方法により、残基の存在する部分の始めと終わりの欄番号をＮ端から順に記述すると、表３のアラインメントは、

Ａ配列：　３，１３，１５，４０

Ｂ配列：　１，２７，３２，４１

Ｃ配列：　２，２６，３０，４０

と記述できる。Ａ配列では欄番号３から１３まで、１５番から４０番までの欄に残基が存在し、Ｂ配列では、欄番号１番から２７番まで、３２番から４１番までの欄に残基が存在することを示す。Ｃ配列についても同様である。

【００３３】

また、各配列につき、各ギャップ部分と残基の存在する部分の残基数を交互に示すことにより、特定の配列を基準とせず、また欄番号によらずに、ギャップ情報を記述することができる（方法③）。この方法により、ギャップの残基数を先に記述した場合、表３のアラインメントは、

Ａ配列：　２，１１，１，２６，１

Ｂ配列：　０，２７，４、１０

Ｃ配列：　１，２５，３，１１，１

と記述できる。Ａ配列のギャップ情報中２、１，１は各ギャップの残基数で、その間に連続した１１残基と２６残基のアミノ酸残基が置かれることを示している。Ｂ配列については、Ｎ端からの２７残基に続き、４残基分のギャップを挟んで、１０残基あることを示し、Ｃ配列については、Ｎ端の１残基のギャップに続いて２５残基、３残基分のギャップに続いて１１残基があり、最後に１残基のギャップがあることを示している。

【００３４】

＜通常の表現形式のアラインメント情報への変換＞

どの方法で記述したギャップ情報でも、通常の表現形式のアラインメント情報に変換できる。まず、ギャップを含む全残基を列記するのに必要な欄の数を計算し、用意する必要がある。ギャップ情報にしたがって、その各欄に各配列の残基番号またはハイフンを対応させれば表３の形式のアラインメント情報が得られ、さらに各配列の残基番号に対応したアミノ酸残基を当てはめれば、通常の表現形式のアラインメント情報が再現できる。

【００３５】

表３のアラインメント情報は、以下のようにして再現できる。方法②による表２のアラインメントに対するギャップ情報（Ａ配列「３，１３，１５，４０」、Ｂ配列「１，２７，３２，４１」、Ｃ配列「２，２６，３０，４０」）から、必要な欄の数は、最大の欄番号と同じ４１である。４１の欄を用意し、Ａ配列については全３７残基をＮ端から順に、欄番号で３番から１３番，１５番から４０番の各欄に並べる。Ｂ配列については全３７残基をＮ端から順に、欄番号で１番から２７番及び３２番から４１番の各欄に並べればよく、Ｃ配列については全３６残基を、欄番号で２番から２６番及び３０番から４０番の各欄に並べればよい。

【００３６】

方法③による表３のアラインメントに対するギャップ情報（Ａ配列「２，１１，１，２６，１」、Ｂ配列「０，２７，４、１０」、Ｃ配列「１，２５，３，１１

，１」）から、必要な欄の数は、Ａ配列の残基数を全部加えて、２＋１１＋１＋２６＋１＝４１と計算される。Ｂ配列から計算しても（２７＋４＋１０＝４１）、Ｃ配列から計算しても（１＋２５＋３＋１１＋１＝４１）、同じ数になる。この欄に対して、Ａ配列についてはＮ端から順に、２残基のギャップの後に１１残基、１残基のギャップの後に２６残基並べ、最後に１残基のギャップを並べればよい。

Ｂ配列についてはＮ端から順に２７残基並べ、４残基分のギャップの後に１０残基を並べればよい。Ｃ配列についても同様である。

【００３７】

一般に、通常の表現形式のアラインメント情報は、含まれる配列群の一部を削除したり、他の配列を加えることによって変化する。これはギャップの入り方が変わるためである。このため、既に利用可能なアラインメント結果のうちの一部を用いる場合や、他の配列を加えてアラインメントする場合には、注意深くアラインメントを修正する必要があり煩雑である。本発明のギャップ情報の記述方法は、含まれる配列群の変更がギャップ情報間の演算を通して容易に行えるのが特徴である。

【００３８】

＜アラインメント情報からの配列の抽出＞

表３のアラインメント情報から、その一部の配列、例えばＢを除去して、ＡとＣのアラインメント情報を通常の表現形式で取り出す場合の手順を示す。方法②による表３のアラインメントに対するギャップ情報（Ａ配列「３，１３，１５，４０」、Ｂ配列「１，２７，３２，４１」、Ｃ配列「２，２６，３０，４０」）から、Ａ配列とＣ配列の情報（「３，１３，１５，４０」、「２，２６，３０，４０」）を取り出す。１から４１欄の間で、両方の配列でギャップになっている欄番号（この場合、１と４１）を演算的に探し、これらの欄番号を左側（欄番号の小さい方）に詰める。その結果、欄の数は３９となり、表５のアラインメントが得られる。

【００３９】

【表５】

R	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	-	1	2	3	4	5	6	7	8	9	10	11	-	12	13	14	15	16	17	18	19	20
C	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

R	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
A	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
C	23	24	25	-	-	-	26	27	28	29	30	31	32	33	34	35	36

【００４０】

方法③による表３のアラインメントに対するギャップ情報（Ａ配列「２，１１，１，２６，１」、Ｂ配列「０，２７，４、１０」、Ｃ配列「１，２５，３，１１，１」）については、方法②によるギャップ情報に変換後、上記と同様に配列の抽出がギャップ情報の演算により容易に行える。

【００４１】

Ａ配列について、方法③のギャップ情報から方法②のギャップ情報への変換の例を以下に示す。方法③のギャップ情報「２，１１，１，２６，１」は、ギャップ部分と残基の存在する部分の残基数を交互に記述したものであり、残基の存在する部分は２カ所ある。その各部分の欄番号の始めと終わりは、２＋１＝３，２＋１１＝１３　及び　２＋１１＋１＋１＝１５，２＋１１＋１＋２６＝４０と計算できるので、方法②によるギャップ情報「３，１３，１５，４０」に変換できる。Ｂ配列についても同様で、方法③のギャップ情報「０，２７，４、１０」は、０＋１＝１，０＋２７＝２７，０＋２７＋４＋１＝３２，０＋２７＋４＋１０＝４１の演算によって、方法②によるギャップ情報「１，２７，３２，４１」に変換できる。

【００４２】

＜複数のアラインメント情報の統合＞

本発明の方法によれば、表６のような共通の配列が存在する２以上のアラインメント結果をギャップ情報の演算により容易に統合することができる。

【００４３】

【表６】

アラインメント１

A － － 1 2 3 4 5 6 7 8 9 10 11 － 12 13 14 15 16 17 18 19 20

B 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

C － 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

A 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 －

B 24 25 26 27 － － － － 28 29 30 31 32 33 34 35 36 37

C 23 24 25 － － － 26 27 28 29 30 31 32 33 34 35 36 －

アラインメント２

A － － － 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

D 1 2 3 4 5 6 7 8 9 10 11 12 13 － 14 15 16 17 18 19 20 21 22

A 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 －

D 23 24 25 － － 26 27 28 29 30 31 32 33 34 35 36 37 －

【００４４】

アラインメント１についての方法②によるギャップ情報は、Ａ配列「３，１３，１５，４０」、Ｂ配列「１，２７，３２，４１」、Ｃ配列「２，２６，３０，４０」であり、アラインメント２についてのギャップ情報は、Ａ配列「４，４０」、Ｄ配列「１，１３，１５、２６，29，４０」である。共通なＡ配列のギャップ情報から、アラインメント１については、Ｎ端に１残基の新たなギャップを、アラインメント２については、欄番号１４と１５の間に新たなギャップを１残基分を設ける必要があることが演算からわかる。

【００４５】

そこで、アラインメント１に含まれる全配列で、Ｎ端のギャップのために欄番号を１つずつ大きくした結果、ギャップ情報は、Ａ配列「４，１４，１６，４１」、Ｂ配列「２，２８，３３，４２」、Ｃ配列「３，２７，３１，４１」となる。

アラインメント２では、欄番号の１４と１５の間の新たギャップの導入のため、ギャップ情報は、Ａ配列「４，１４，１６、４１」、Ｄ配列「１，１３，１６、２７，３０，４１」となる。このように両アラインメントにおいて、Ａ配列のギャップ情報が同一になればよい。これらの情報を、上記の手順に従って通常の表現形式に変換すると、表７のように統合されたアラインメントが得られる。統合されたアラインメントでの必要な欄の数は、含まれる最大の欄番号（Ｂ配列）から４２である。

【００４６】

【表７】

R	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	-	-	-	1	2	3	4	5	6	7	8	9	10	11	-	12	13	14	15	16	17	18
B	-	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
C	-	-	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
D	1	2	3	4	5	6	7	8	9	10	11	12	13	-	-	14	15	16	17	18	19	20

R	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
A	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	-
B	22	23	24	25	26	27	-	-	-	-	28	29	30	31	32	33	34	35	36	37
C	21	22	23	24	25	-	-	-	26	27	28	29	30	31	32	33	34	35	36	-
D	21	22	23	24	25	-	-	26	27	28	29	30	31	32	33	34	35	36	37	-

【００４７】

方法③によるギャップ情報についても、方法②によるギャップ情報に変換することによって、上記と同様に演算的に統合が可能である。共通の配列がない場合には、両アラインメントから、いずれかの配列が共通となるようそれぞれのアラインメントから１ヶづつの配列を選んでアラインメントを行う。

【００４８】

本発明の方法において、ギャップ情報を表す数字は、コンピュータで効率よく処理できるようにバイト単位で表現するとよい。例えば、１バイト以内で表現でき

る数字は１バイトで表現し、１バイトで表現できないときは、複数バイトで表現することができる。何バイトで表現されているかを表すために、特定のビットにフラグをたててもよい。また、数字のプラス、マイナス符号のかわりに、データの中に符号ビットを設けても良い。また、単純なアルゴリズムにより、これらの数字群を別データに変換したものをギャップ情報として利用することもできる。

【００４９】

ギャップ情報をデータベースに保存するにあたり、配列を特定する記号又は番号（配列ID）を付する必要がある。また、アラインメント情報ごとにそれを特定する記号又は番号（アラインメントID）をつけ、そのIDから配列IDを検索できるようにしておくことも可能である。同じアラインメントIDに属する一群のギャップ情報は、その区切りがわかるように連結して保存することも可能である。配列情報はギャップ情報を含むアラインメント情報のファイル又はデータベース中に保存されている必要はない。ギャップ情報を保存した媒体と同一媒体中に配列情報を保存することが管理上望ましいが、検索に利用可能な他の媒体中に保存されていてもよい。

【００５０】

また、他のデータベース中に保存されている配列情報を利用可能な場合には、ギャップ情報のみをデータベースに保存すればよい。使用頻度の高い配列情報はギャップ情報とともに保存し、他は外部データベースを利用することも可能である。また、データベースに含める配列情報には、配列ID、蛋白質名、アミノ酸残基数、アミノ酸配列などの情報のほか、該蛋白質の存在する生物種や組織、サブタイプなどの情報も含めてもよい。あるいは、これらの情報を関係づけられたテーブルで別々に管理してもよい。

【００５１】

＜通信プロトコル＞

アラインメント情報を配列情報とギャップ情報とに分離した形で送信側から受信側に送信し、受信側でアラインメント情報を再構成することにより、効率的にアラインメント情報を通信することができる。まず、上記に説明した方法に従ってアラインメント情報を配列情報とギャップ情報とに分離する。ギャップ情報が対

応するアミノ酸配列を一意に特定できる配列 ID をギャップ情報に付加し、これを送信側から受信側に送る。受信側のデータベースに配列 ID が対応するアミノ酸配列情報がある場合にはそれを利用し、受信側にデータがない場合には送信側に依頼して配列 ID に対応する配列情報を送信させるか、別途、利用可能な他のデータベースから配列IDに対応する配列情報を入手する。受信側では、上記に説明した方法に従って、ギャップ情報からアラインメント情報を再構成することができる。

【００５２】

また、別の方法としては、まずアラインメント情報を配列情報とギャップ情報とに分離する。ギャップ情報が対応するアミノ酸配列を一意に特定できる配列 ID をギャップ情報に付加し、これを送信側から受信側に送る。また配列情報では重複がないようにしてギャップ情報で対応しているIDの配列のみを送信側から受信側に送る。この際、ギャップ情報と配列情報が分離されてさえいれば、送信の順序は関係ない。

【００５３】

【実施例】

以下、本発明を実施例によりさらに具体的に説明するが、本発明の範囲は下記の実施例に限定されることはない。以下の実施例においては、本発明の好ましい態様として方法③によって提示されたギャップ情報を用いたが、上記に例示した他の方法やさらに別な方法によっても、アラインメント情報をギャップ情報と配列情報に分けて扱うことができることはいうまでもない。

【００５４】

例１

４個のアミノ酸配列のアラインメント情報を表８に示すようにギャップ情報と配列情報にわけ、ギャップ情報をデータベースに保存した。表中、各アミノ酸配列は、配列情報を取り出すために配列IDを付して特定した。ギャップ情報のうち、配列ID=000001は基準配列を表し、配列ID = 000002以下については、基準配列に対するギャップ情報である。また、配列IDに対応したアミノ酸配列情報において、各配列には同一の配列IDを付した。

【００５５】

【表８】

（ギャップ情報）

配列 ID	ギャップ情報
000001	3，11，1，26，1
000002	1，27，4，10，0
000003	2，25，3，11，1
000004	0，13，2，12，2，12，1

（配列情報）

配列 ID	蛋白質名	アミノ酸残基数	アミノ酸配列
000001	xxxxxxxx	37	MISLIAALAVDARVIGMENAMPWNLPADLAFKRNTLD
000002	xxxxxxxx	36	VKMISLIAALAVDRVIGMENAMPWNLPAFKAERNTL
000003	xxxxxxxx	36	AMISLIAALAVDRVIGMENAMPWNLPAWFKRNTLDV
000004	xxxxxxxx	37	SEAMISLIAALAVDRVIGMENAMPWNLPADLAWFKRNTLD

【００５６】

例２

表９のアラインメント（甲）の縦列の属性情報を表１０のアラインメント（乙）に統合して印付けした。

【００５７】

【表９】

アラインメント情報（甲）

縦列属性情報	--*-*---*----#-#***--#*-****-----****-------
Sequence A	--MISLIAALAVD-VIMGRHTWESIVYEQFLPKAQHDLYIA--
Sequence B	RSMLSIVAVCQNDAVIMGKKTWFSIVY----AKAQHEKFVSPA

【００５８】

【表１０】

アラインメント情報（乙）

Sequence B　　-RSMLSIVAVCQN---DAVIMGKKTWFSIVYAKAQHEKFVSPA

Sequence C　　A-SVVSLAAVCRNNKPEAVLMMKKSWFSLLYAKAQHEKFVSPV

【００５９】

表９では Sequence A と Sequence B において縦列の対応においてアミノ酸配列が一致している箇所を縦列属性情報として＊で示してある。また機能上重要なアミノ酸の箇所を＃で示してある。表８のようにアラインメント情報を配列情報とギャップ情報に分離したやり方と全く同じ手順を用いて、表９の縦列属性情報において「-」をギャップとみたてることで、縦列属性情報を表11のように属性種類情報と縦列位置情報に分離した。この場合の縦列位置情報は方法③のギャップ情報の表現と同じである。

【００６０】

【表１１】

属性種類情報　＊＊＊＃＃＊＊＊＊＃＊＊＊＊＊＊＊＊＊＊

縦列位置情報　2, 1, 1, 1, 2, 1, 4, 1, 1, 4, 2, 2, 1, 4, 5, 4, 7

【００６１】

表１１の縦列位置情報と、表９の Sequence B のギャップ情報と、表10のSequence Bのギャップ情報から表10のアラインメント（乙）における縦列位置情報を計算したのが表12である。

【００６２】

【表１２】

縦列位置情報　3, 1, 1, 1, 2, 1, 7, 1, 1, 4, 2, 2, 1, 4, 1, 4, 7

【００６３】

表12の縦列位置情報と、表11の属性種類情報からアラインメント（乙）上での縦列属性情報を示したのが表13である。表９と表13を見比べて明らかなようにアラ

インメント（甲）とアラインメント（乙）に共通な Sequence B 上で縦列属性情報の対応がとれている。

【００６４】

【表１３】

アラインメント情報（甲）の縦列情報を Sequence B 上で対応するようにしてマップしたアラインメント情報（乙）

甲の縦列属性情報	---*-*--*--------#-#****--#*-****-****--------
Sequence B	-RSMLSIVAVCQN---DAVIMGKKTWFSIVYAKAQHEKFVSPA
Sequence C	A-SVVSLAAVCRNNKPEAVLMMKKSWFSLLYAKAQHEKFVSPV

【００６５】

【発明の効果】

本方法によれば、アラインメント情報を保存する際に、配列情報が重複することがなく、ギャップ情報も数個の数字でデータ化されるため、全体として極めて少ないデータ量で保存することができる。また、それらの情報から簡単に通常の表現形式のアラインメント情報も取り出すことができる。さらに、対応付け情報をそれ自体の間で演算することにより、アラインメント情報に含まれる配列群の編集や統合したアラインメント情報の取り出しが可能になり、アラインメント情報の再利用以外にも多様な応用が可能である。従って、本発明の方法により、データベースや各種記録媒体（例えば、磁気記録媒体や光記録媒体など）へのアラインメント情報の保存効率が飛躍的に高まり、大量のアラインメント情報を蓄積し、それらの再利用が容易なデータベースをより有効に作成することが可能になる。

【００６６】

また、配列情報とギャップ情報とを分離して管理できるため、データベースの整合性及び保守性を保つことが容易になる。特に、リレーショナルデータベースにおいては、より正規化された状態でデータが扱えるため、データベース利用の可能性がさらに高まる。さらに、本発明の方法に従ってアラインメント情報を送信する場合には、受信側にすでにある配列情報を送信せずに済み、通信効率が向上するとともに、受信側でアミノ酸配列情報の重複が生じない。特に膨大な量のア

ラインメント情報を送信する場合のほか、データベースの複製を通信を通して作成する場合や、クライアント-サーバーシステムの間でアラインメント情報をやりとりする場合などに有効である。

【書類名】　　要約書

【要約】

【課題】　アミノ酸残基をできる限り一致するように複数の配列を並置したアラインメントの情報を効率的に記述又は保存する方法を提供する。

【解決手段】　アミノ酸配列又は核酸配列のアラインメント情報を配列情報と配列間の対応関係を表すギャップ情報とに分離して記録する方法、又はこれらの情報のうち少なくともギャップ情報を記録媒体に保存することを特徴とするアラインメント情報の保存方法。

出　願　人　履　歴　情　報

識別番号　　〔５９７０５１１４８〕

１．変更年月日　　１９９７年　４月１１日
［変更理由］　新規登録
住　所　　東京都文京区本郷５丁目２４番５号　角川本郷ビル４Ｆ
氏　名　　株式会社医薬分子設計研究所